OPEN ACCESS

KING'S College LONDON

King's Research Portal

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Batlajery, B. V., Weal, M., Chapman, A., & Moreau, L. (Accepted/In press). Belief Propagation Through Provenance Graphs. In *IPAW'2018: 7th International Provenance and Annotation Workshop* Springer-Verlag Berlin Heidelberg.

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

BELIEF PROPAGATION THROUGH PROVENANCE GRAPHS

Belfrit Victor Batlajery[1], Mark Weal[1], Adriane Chapman[1], and Luc Moreau[2]

[1] University of Southampton
email:{b.v.batlajery,m.weal,adriane.chapman}@soton.ac.uk
[2] King's College London
email:luc.moreau@kcl.ac.uk

Abstract. Provenance of food describes food, the processes in food transformation, and the food operators from the source to consumption; modelling the history food. In processing food, the risk of contamination increases if food is treated inappropriately. Therefore, identifying critical processes and applying suitable prevention actions are necessary to measure the risk; known as due diligence. To achieve due diligence, food provenance can be used to analyse the risk of contamination in order to find the best place to sample food. Indeed, it supports building rationale over food-related activities because it describes the details about food during its lifetime. However, many food risk models only rely on simulation with little notion of provenance of food. Incorporating the risk model with food provenance through our framework, *prFrame*, is our first contribution. *prFrame* uses Belief Propagation (BP) over the provenance graph for automatically measuring the risk of contamination. As BP works efficiently in a factor graph, our next contribution is the conversion of the provenance graph into the factor graph. Finally, an evaluation of the accuracy of the inference by BP is our last contribution.

1 Introduction

Provenance of food is well understood by both business and the public. Notions of *Appellation d'Origine Contrôllée* are regulatory labels indicating that some food products can be trusted to originate from a given region, thus vouching for the authenticity and quality of the products. Likewise, organic labels encompass more or less stringent guarantees that adequate processes have been followed in the production of food products. The provenance model PROV (PROV-DM) complemented by domain-specific ontologies [1][2] have been used to describe processes of the food supply chain, enabling such descriptions to be shared and queries over them to be answered. These capabilities allow confidence in food products and processes to increase. For instance, the requirement for food operators to identify suppliers one level up and customers one level down can easily be addressed using provenance-based modelling of the food supply chain [3].

Regulations demand that food operators undertake due diligence [4]. While this term is not formally defined in law, it is usually understood to include

identifying all the food safety critical stages of food production, storage and distribution, then identifying suitable control measures to adequately prevent the risk of food safety failures and putting in place appropriate management control procedures to ensure they effectively happen [4][5].

Our claim in this paper is that provenance models such as PROV can be the basis for food operators to develop a rationale for control procedures. Indeed, our discussions with them show that food samples are analysed to check contamination levels as part of a due diligence process to manage risk. However, such samplings are costly in terms of resources, and a rationale needs to be developed on how best to sample food supply chains. Regulators and food operators are constantly on the lookout for better ways to measure, track, and analyse risk in the food supply chain. In this paper, we discuss two techniques, adapted to operate over provenance graphs, which result in a powerful tool to reason, estimate, and understand risk of contamination across the food supply chain, over which we have partial knowledge of level of contamination.

First, PROV provenance can be used to model the food supply chain in the Modular Process Risk Model (MPRM), which is a tool for Quantitative Microbial Risk Assessment (QMRA) [6]. MPRM uses Monte-Carlo (MC) simulation, which is a computer-based technique allowing variation of randomly distributed inputs to be propagated through mathematical models [7], to generate bacterial concentration with the aim to understand the distribution of bacteria in the food supply chain. This approach relies on the directed nature of provenance graphs, and propagate bacterial concentration along edges of these graphs, according to evidenced formulae of micro-organisms transmissions. However, MPRM does not support any actual knowledge of contamination level as it relies on distributions of bacterial concentration, derived from past studies.

Second, in this context, Belief Propagation (BP) is a technique that takes observations of contamination levels in the food supply chain to calculate the marginal distribution for each unobserved node, conditional on these observed nodes [8]. BP, initially defined by Pearl, has been showed to operate on trees, but also to provide useful approximations for graphs. It requires a notion of Factor Graph (a bipartite graph containing nodes for variables and factors), which we demonstrate can be easily derived from provenance graphs.

The aim of this paper is to introduce a *prFrame*, as a framework to estimate risk of contamination in a food supply chain described by provenance, which allows for observations (by directly sampled contamination levels) to be taken into account, as well as estimates to be inferred for unobserved part of the chain. We demonstrate the effectiveness of the methodology within this framework lies when new evidence (i.e. sampling report) can easily be incorporated to more accurately estimate the actual risk.

The concrete contributions of the paper are as follows:

1. A Monte-Carlo based simulation technique to derive contamination levels of provenance-based descriptions of a food supply chain.
2. A transformation of food supply chain provenance graphs into factor graphs to enable sum-product algorithm as a variant of Belief Propagation.

3. An evaluation framework, allowing contamination levels to be systematically hidden in a provenance described supply chain (effectively creating unobserved nodes) to generate estimations of contamination levels through Belief Propagation.

Following this introduction, background to support our work is given in section 2. We present our case study and our approach with *prFrame* in section 3 and 4 subsequently. Section 5 shows how our approach can be applied and section 6 concludes the paper and suggests potential future work.

2 Background

In this section, some theoretical concepts are presented. In general, we have two intersected concepts, namely provenance to describe what happened to food and BP to infer the risk of contamination over the provenance graph.

2.1 Provenance

The World Wide Web Consortium (W3C) defines provenance as *a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing in the world* [9]. It contains the description of data and the processes involved during the data lifetime, such as how something is derived, who is responsible for certain actions, what the consequences and the risks of an activity are, etc. As provenance describes the lifetime of something, it can provide a crucial information for investigation.

The Provenance Data Model (`PROV-DM`) and Provenance Ontology (`PROV-O`) [9][10] enable the modelling of something in provenance. As `PROV-O` is designed to be domain agnostic, it often gets extended in specific domains. For instance, Markovic et al. extend the `PROV-O` to monitor food safety by documenting constraints that may be associated with an HACCP plan [2] and Batlajery et al. provide *prFood* ontology to capture and model food and risk along the food supply chain [1]. Another works by Ali and Moreau [11], Packer et al. [12], and Markovic et al. [13] also extend `PROV-O` for their specific purposes.

2.2 Belief Propagation

BP is an approach to perform inference based on message passing algorithm. Here, we focus on sum-product, as an algorithm in the BP family.

Theory of belief propagation In Probabilistic Graphical Model (PGM), probability theory and graph theory are utilised to capture the knowledge in graph-based representations [14]. Probability is about measuring uncertainty of an occurrence in the world, which refers to the degree of confidence that an event will occur [15]. For example, the probability $P(X)$ of an event X quantifies the degree of confidence that X will occur. With $P(X)$=1, we are certain that one

of the outcomes in X occurs and $P(X)$=0 indicates that all outcomes in X are impossible. Other probability values between 0 and 1 represent options that lie between them. Probability can be expressed in 2 fundamental rules, sum rule and product rule, which become the basic calculations of sum-product algorithm.

$$(a)\ \textbf{sum rule} \quad P(X) = \sum_{Y} P(X,Y) \qquad (b)\ \textbf{product rule} \quad P(X,Y) = P(Y|X)P(X) \tag{1}$$

In Equation 1, $P(X)$ is referred to as marginal distribution over the distribution of random node X and is simply verbalized as the probability distribution of X. In many cases, the questions often involve the values of several random nodes or a Joint Probability Distribution (JPD), written as $P(X,Y)$. Similarly, a Conditional Probability Distribution (CPD) can be verbalized as the probability of Y given X or $P(Y|X)$ that specifies the belief in Y under the assumption that X is known (observed) with certainty [16]. Entering an evidence to update our belief about the probability is often mentioned as propagation and its mechanism with BP is described in the following paragraph.

Mechanics behind belief propagation BP relies on an iterative message passing algorithm inherently from bayesian procedure to perform an inference efficiently. This technique explores the conditional independence relationship over a Factor Graph. A factor graph is a bipartite graph that expresses the global function into a product of local functions [18]. This graph consists of 2 types of nodes, namely a variable node for each node in the network and a factor node for each factor *f(x)* in the joint distribution between nodes.

The message passing algorithm allows the nodes to communicate their local state by sending messages over the edges [14][19][20]. By local, we mean that a given node updates the outgoing messages on the basis of incoming ones from the previous iterations. In general, the messages are passed around and get updated until a stable belief state is reached (convergence). However, depending on the type of graph, some may not reach the convergence due to circular reasoning. In the context of food provenance, the circular process exists in the event of splitting and joining food. Thus, having the provenance of food with a tree-based structured, such as the sequential linear chain from the source to consumption, can guarantee the convergence. Although convergence is not guaranteed, BP has been found to have outstanding empirical success in loopy graphs too [14].

3 Food supply chain as a use case

This section presents a case study where a notion of provenance is needed in the food domain in order to achieve due diligence. The study involves a food risk model for the food domain and its modelling by provenance.

3.1 Food provenance and food regulations

In order to achieve due diligence, food regulations (e.g. ISO 9000, Food Safety Act (FSA) 1990, HACCP, etc.) are created for assuring food that people consume

is safe [4][21]. By identifying what, where, when, who, and how food is handled, regulators and food operators can have an overview of potential contamination and have more comprehensive way of understanding the risk.

Definition 1. *Food Provenance is a record that describes a food product and its ingredients; the processes involved in food transformation; and food operators who are responsible for those processes in the food supply chain.*

In modelling food provenance into the standardized provenance format, we use PROV because of its capability to capture and describe the entities, activities, and agents that may have influenced the piece of data about food. The modelling is performed by codifying a food stage (e.g. *Prepared*, *Cooked*, etc.) as *prov:Entity*, a food process (e.g. *preparing*, *cooking*, etc.) as *prov:Activity*, and a food operator as *prov:Agent*. Figure 1 shows an excerpt of food provenance.

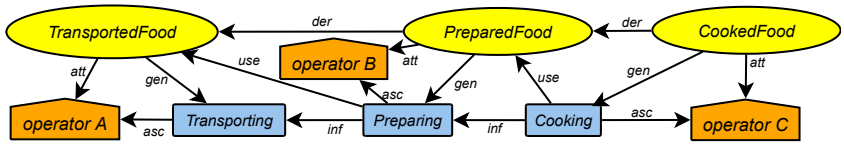


Fig. 1. An excerpt of provenance graph of the food supply chain.

Figure 1 describes the provenance of food, which conveys its history and the information about risk of contamination. To model the risk, the *prFood* ontology [1] is used to capture the necessary data for risk calculation, such as bacterial concentration, contamination level, risk factor, etc as attributes of entities.

Definition 2. *Bacterial Concentration is the total bacteria in food.*

Definition 3. *Contamination Level is a range of values to categorize bacterial concentration.*

Definition 4. *Risk Factor is any aspect that contributes to the risk of contamination, such as improper storage, time and temperature abuse of food, etc. [22].*

3.2 Modular Process Risk Model (MPRM)

MPRM is a process-driven framework to estimate the risk of food contamination based on how food is handled [6]. This framework splits the food supply chain into smaller modules and the transmission of bacteria is calculated based on the well-known formulae with the MC simulation. The simulation selects a random value from the distribution of a risk factor to generate bacterial concentration after each food process, and it will be the input for the next process.

MPRM supports 6 basic processes that can affect the bacterial concentration after the food process. They are Growth, Inactivation, Partitioning, Mixing, Removal, and Cross Contamination. Growth and inactivation are two basic microbial processes, which are strongly depending on the characteristic of bacteria investigated and the surrounded environmental condition. Partitioning, mixing,

removal, and cross-contamination are 4 handling processes. Partitioning occurs when a major unit of food is split up into several minor units, while mixing describes the opposite process. Removal is a process where some units are removed and cross-contamination describes the transmission of bacteria between objects.

4 The prFrame Framework

This section discusses *prFrame*, our proposed framework that incorporates Provenance, Risk Model, and PGM to achieve due diligence. With multiple food risk models that use a MC simulation, the input-output interaction of bacterial concentration only works in one direction (forward, from source to destination), making predicting the contamination level before a food process difficult, given the bacterial concentration after that process. In addition, its capability to incorporate an actual bacterial concentration is limited. Meanwhile, BP is a non-directional approach as it propagates information forward and backward. Thus, inferencing is easier with additional observed information anywhere in the chain. The pseudocode of *prFrame* is shown in Algorithm 1 and is described below.

Algorithm 1: prFrame Algorithm

Input : pG: Provenance Graph
Output: $infBin$: Inferred Bacterial Level

var $bConc$: Bacterial Concentration ;
var $preBin$: Predicted Bacterial Level ;
$\langle bConc, preBin\rangle \leftarrow \underline{monteCarlo}(pG)$;
;
var $preBin$: Predicted Bacterial Level ;
var $binMtx$: Bin Matrix ;
$binMtx \leftarrow \underline{computeBinMtx}(preBin)$;
;
var $binMtx$: Bin Matrix ;
var $jpdMtx$: JPD Matrix ;
var $cpdMtx$: CPD Matrix ;
$\langle jpdMtx, cpdMtx\rangle \leftarrow \underline{computeCpd}(binMtx)$;
;
var pG: Provenance Graph ;
var $cpdMtx$: CPD Matrix ;
var $pGcpd$: Provenance Graph with CPD;
$pGcpd \leftarrow \underline{attachCpd}(pG, cpdMtx)$;
;
var $pGcpd$: Provenance Graph with CPD;
var fG: Factor Graph ;
$fG \leftarrow \underline{convertPG}(pGcpd)$;
;
var fG: Factor Graph ;
var e: Observed nodes ;
var i: Inferred nodes ;
var $infBin$: Inferred Bacterial Level ;
$infBin \leftarrow \underline{beliefPro}(fG, e, i)$;

4.1 Food risk model with monte-carlo simulation

Our framework begins with a given provenance graph that describes food. The provenance graph is expected to hold data about risk factors as parameters to simulate the flow of food based on MPRM. An MPRM basic process in a food process depends on the activities described and the assumption hold in that food process. For example, it is assumed that the number of microbes increases during the transporting process; hence, growth model becomes the basic process for transporting. Changing or adding a basic process will affect the formula to predict the number of microbes, which is not the scope of this paper. We refers

the readers to [22] for the details of risk factors and their distributions in each food process as well as the formula for each MPRM basic process.

The simulation is needed as we do not know the exact risk factors, such as time and temperature in processing food, leading us to only have partial information about contamination levels. With this reason, we estimate bacterial concentration by conducting MC simulation, which takes into account all the possible values of risk factors in form of a distribution, to predict contamination level along the provenance network. The MC simulation is performed the same as in [22], which generates predicted bacterial concentration after each food process.

Each generated bacterial concentration is categorized into the contamination level. The aim for categorization is that it is easier to compare the actual data with the categorical data (contamination level) rather than with the continuous data (bacterial concentration) in order to infer the updated risk of contamination. Thus, each contamination level counts food that have bacterial concentration within its defined range (Alg.1 *line 3*). In the end, a *Bin matrix* is constructed to capture all possible combinations between contamination levels before (upwards) and after (downwards). The column and row of the matrix represent the levels upward and downward consecutively (Alg.1 *line 7*). For example, Figure 2 shows that there are total 24 food that in the transporting process (becomes Transported Food) and storing process (becomes Stored Food). Four of them had microbial level 1 after transporting and level 2 after storing.

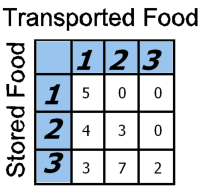


Fig. 2. An example of a bin matrix. A blue square represents the level of contamination.

4.2 Belief propagation in the provenance network

A Joint Probability Distribution (JPD) is captured in a *JPD matrix* by dividing each value in the Bin matrix with the total number in Bin matrix (total food used that have undergone the food process). Subsequently, a *CPD matrix* is derived by dividing each value of the JPD matrix with its corresponding row as the row represents the level downward the food process (Alg.1 *line 12*). A complete bin matrix, jpd matrix, and cpd matrix are presented in on-line appendix (https://goo.gl/hXvici). Next, the CPD matrix is added as an attribute in the provenance graph (Alg.1 *line 17*) and the conversion into a factor graph is performed (Alg.1 *line 21*).

In a factor graph, a factor can be described as a function that takes arguments from the random nodes and return a value for every possible combinations over those random nodes. A CPD is used as a factor, which holds the notion of conditional probability for every *prov:Entity* that is linked with a *prov:Activity*

via both *prov:usage* (*use*) and *prov:wasGeneratedBy* (*gen*), in the present of a *prov:wasDerivedFrom* (*der*) that identifies the origin and the result of the food process for a CPD matrix. Algorithm 2 shows the pseudocode of the conversion.

Algorithm 2: function factorGraph(*pGcpd*)

```
   Input  : pGcpd: Provenance Graph with CPD
   Output: fG: Factor Graph
 1 var n_x: Variable node ;
 2 var f_x: Factor node ;
 3 var unEdge_x: Undirected Edge ;
 4 var o: Object ;
 5 foreach o ∈ pGcpd do
 6     if type(o)=prov:Entity AND type(o)=prFood:FoodStage then
 7         n_x ← convertEntity(o) ;
 8     end
 9     if type(o)=prov:Activity AND type(o)=prFood:FoodProcessing then
10         f_x ← convertActivity(o) ;
11     end
12     if type(o)=prov:usage OR type(o)=prov:wasGeneratedBy then
13         unEdge_x ← convertEdge(o) ;
14     end
15 end
16 return Factor Graph (fG)
```

Overall, the conversion maps each *prov:Entity* into a variable node ($\mathbf{n_1}$,...,$\mathbf{n_x}$) (Alg.2 *line 6*) and each *prov:Activity* into a factor node $\mathbf{f_x}$ (Alg.2 *line 9*) in the factor graph. Only a *prov:Activity* that has the type *prFood:FoodProcessing* will be converted into a factor node, and a *prov:Entity* of type *prFood:FoodStage* will be converted to a variable node. The factor node $\mathbf{f_x}$ holds the notion of CPD, which is a factor to determine the probability of each variable nodes that are connected to it ($\mathbf{n_1}$,...,$\mathbf{n_x}$). In the conversion, we ignore *prov:Agent* to make the graph as simple as possible. Figure 3 shows an example of the conversion.

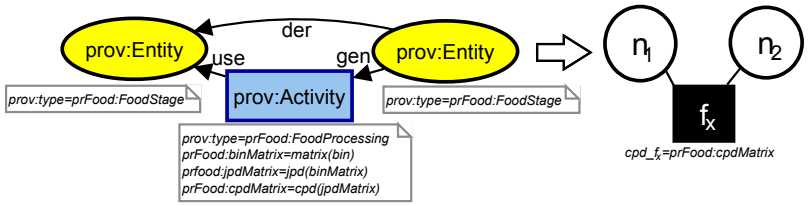


Fig. 3. A conversion from provenance graph into factor graph.

In Figure 3, in order to link the factor nodes with the variable nodes, we identify *prov:wasGeneratedBy* (*gen*) and *prov:usage* (*use*) and convert them into undirected edges (Alg.2 *line 12*) provided a corresponding *prov:wasDerivedFrom* (*der*) exists (as its notion has been encapsulated in the CPD matrix). For example, the probability of $\mathbf{n_x}$ given $\mathbf{n_{x-1}}$ has implied the derivation between $\mathbf{n_x}$ and $\mathbf{n_{x-1}}$. Finally, the sum-product algorithm that utilizes bayesian rules is applied

to calculate the likelihood of a certain event (Alg.1 *line 27*). The figures of initial provenance graph and factor graph are available in on-line appendix.

4.3 Methodology to infer risk of contamination

As a framework, *prFrame* is intended to automatically infer the risk of food contamination. It incorporates the general food risk model that uses MC simulation, MPRM, with the inference technique, BP. BP infers the actual contamination level by propagating belief based on the previous knowledge and the actual data (i.e. sampling result). Our methodology compares the inference of contamination level by BP ($InfBin$: Inferred Bacterial Level) with the prediction by the MC simulation ($prBin$: Predicted Bacterial Level).

The aim in this methodology is to understand the accuracy of inference across the food provenance network by capturing exhaustively experiments, where nodes values are hidden and observed systematically, in order to evaluate the performance of BP. To define an accuracy of inference, consider a bacterial concentration in level 1 that was predicted by the MC simulation. There are three possible inferences by BP. The first inference reveals with 100% probability that the prediction is in level 1. The second inference reveals with 97.6% probability that the prediction is in level 1 and 2.4% probability in level 2. The third inference reveals with 90% probability that the prediction is in level 1 and 5% probability is in both level 0 and level 2. Here, the most accurate inference is the first inference, followed by the second and the third inferences.

5 Evaluation of the methodology

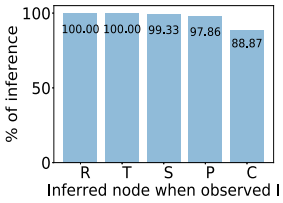
The complete list of chains for our first and second setup is shown in on-line appendix. To keep the calculation and propagation simple, we use an example of a fixed linear network that represents the food chain as configured in [22]. It is also possible to define more complicated network as provenance of food can be non-linear network. However, the use of linear chain in this paper is guaranteed to reach a convergence in inferring with BP for further measurement of accuracy. The defined network comprises 6 food stages, namely *Initial* (**I**), *Retailed*(**R**), *Transported*(**T**), *Stored*(**S**), *Prepared*(**P**), and *Cooked*(**C**). We then predict bacterial concentration and contamination level by performing MC simulation with 50,000 iterations to represent the travelling of food products through this chain and results in 10,502 food being contaminated.

Each bacterial concentration will be categorized into a fixed determined *bin* that represents contamination level. We consider 13 levels of bacteria because the number is precise enough to categorize the bacterial concentration. While adding more levels produces more precise result, it comes with higher computational. We also introduce an *inferred node*, an *unobserved node* that its probabilities are in our investigation when performing an inference. Finally, we perform an inference with BP and compare the results against the prediction of MC simulation.

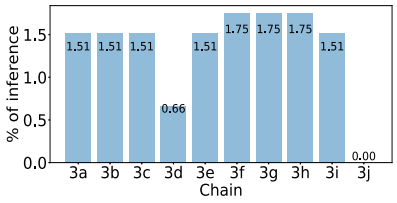
5.1 The effect of the distance and position between nodes

Our first setup intends to measure accuracy of inference based on the distance between an *observed node* and an *inferred node*. We set node **I** as an *observed node* and the remaining nodes will be inferred. Figure 4(a) shows that the accuracy decreases as we infer the further nodes. Inferencing nodes **R** and **T** is always correct as all of the inferences suggest the same level as predicted by MC simulation. The inference becomes less accurate with total 10,432 correct inferences (99.33%) in node **S**. In other words, it can be verbalized as there are 99.33% contaminated food with 88%-100% probability of being correct. Finally, there are 97.86% and 88.87% correct inferences in node **P** and node **C** consecutively. In addition, inferring node **C** by observing node **I**, **R**, **T**, **S**, and **P** for our second setup also suggest the same result.

In regards to the position of nodes, we condition some nodes (solid-filled-node) and let BP does the inference in node **S** as a *inferred node* (dashed-unfilled-node) as shown in several chains (ch.) in the on-line appendix. In Figure 4(b), inferencing node **S** with upward *observed nodes* (nodes **I**, **R**, and **T**) gives the same result (ch.3a, ch.3b, and ch.3c). Among 10,502 inferences, only 159 inferences (1.51%) are with 100% probability of being correct. In fact, more upward *observed nodes* produces the same result too (ch.3e). The inference becomes more accurate if the *inferred node* is set in between the *observed nodes* (ch.3f, ch.3g, and ch.3h) with 1.75% correct inferences. However, the result is less accurate if we observe nodes **I** and **C** (ch.3i) with only 1.51% correct inferences. The opposite result is shown in ch.3d and ch.3j, where a downward node is observed with the remaining nodes unobserved. This scenario shows the deterioration of the accuracy with 0.66% and 0.0% correct inferences consecutively.



(a) Inference with 89%-100% probability of being correct in the first setup.



(b) Inference with 100% probability of being correct in third setup.

Fig. 4. The effect of distance and location of nodes in the accuracy of inferring.

5.2 Analysis of the result

From our evaluation, we conclude that the closer the distance between *observed node* and *inferred node* is, the more accurate the inference will be. This can be proved through the first and second setup. Moreover, the highest accuracy of inference is achieved when the *inferred node* is placed between the *observed nodes*. This is obvious as the upward and downward nodes can infer the middle one with more certainty. In fact, the accuracy is similar when we add more

observed nodes, indicating that the only important nodes are one node upward and one node downward the *inferred node*. Although *inferred node* is located in between *observed nodes*, the accuracy decreases if there is *unobserved node* in between those *observed nodes*, which provokes the uncertainty.

Our evaluation also reveals that observing several nodes prior the inferred one will not improve the accuracy if the nodes downward the *inferred node* remain unobserved. Again, it means that adding more nodes prior the *inferred node* does not affect the inference as long as the downward nodes remaining unobserved. The same result derived if the *observed nodes* are located downward the *inferred node* with remain nodes unobserved. However, the inference is more accurate when observing upward nodes than the downward nodes of the *inferred node*.

6 Conclusion and Future work

We have presented our work on using BP as an inference technique over the provenance network through *prFrame* to infer the conditional probability of a node, given a condition of the others. We conclude that *prFrame* successfully combines BP with the provenance network and our evaluation produces inferences with high accuracy between 89% and 100% of being correct. In the food context, it can be translated as the contaminated food are inferred with 89% to 100% chance of being correct. We believe that more reliable result can be achieved with more sufficient data captured in provenance, such as risk factors or sampling data. From implementation point of view, *prFrame* can accommodate the existence food risk models in order to help food authority achieve due diligence in food.

In a case when a sampling report is used as the actual information, an inference can be performed after the fact that food has travelled to several places as opposed to real time, because sampling analysis can take several days. In this situation, provenance or the past description of food is an important information to explain the reason behind the sampling result and assess the risk to identified the next potential places to sampling food on the basis of the sampling report.

In the paper, we limit our work in a linear network only, while provenance networks are mostly non-linear networks. In fact, many food chains in reality are not a linear chain, such as tree structure. Our investigation reveals that as long as the chain does not have cycle in it, the inference becomes converged. However, even though the food chain have a cycle and the state belief cannot be achieved, an approximate inference with BP has been proven as a good estimation as well.

Finally, in performing the inference, we did not take into account the type of activity of food process to assess the accuracy of inference. We believe that a deeper investigation is required in order to systematically characterise BP-band in inference in provenance trace.

References

1. Batlajery, B.V., Weal, M., Chapman, A., Moreau, L. In: prFood: Ontology principles for provenance and risk in the food domain. IEEE (12 2017)

2. Markovic, M., Edwards, P., Kollingbaum, M., Rowe, A.: Modelling provenance of sensor data for food safety compliance checking. In: International Provenance and Annotation Workshop, Springer (2016) 134–145
3. Thakur, M., Hurburgh, C.R.: Framework for implementing traceability system in the bulk grain supply chain. Journal of Food Engineering **95**(4) (December 2009) 617626
4. Food Standards Agency: Food Law Code of Practice (England)-April 2015. Report, Food Standards Agency (April 2015)
5. Eves, A., Dervisi, P.: Experiences of the implementation and operation of hazard analysis critical control points in the food service sector. International Journal of Hospitality Management **24**(1) (March 2005) 319
6. Nauta, M.J.: A modular process risk model structure for quantitative microbiological risk assessment and its application in an exposure assessment of bacillus cereus in a repfed. RIVM Rapport 149106007 (2001)
7. Duarte, A.S.R.: The interpretation of quantitative microbial data: meeting the demands of quantitative microbiological risk assessment. PhD thesis, National Food Institute, Technical University of Denmark (2013)
8. Pearl, J.: Reverend Bayes on inference engines: A distributed hierarchical approach. AAAI'82. AAAI Press (1982)
9. Moreau, L., Groth, P., Cheney, J., Lebo, T., Miles, S.: The rationale of prov. Web Semantics: Science, Services and Agents on the World Wide Web **35**(4) (December 2015) 235–257
10. Moreau, L., Missier, P.: PROV-DM: The PROV Data Model,. W3C Recommendation REC-prov-dm-20130430, World Wide Web Consortium (April 2013)
11. Moreau, L., Ali, M.: A provenance-based policy control framework for cloud services. (May 2014)
12. Packer, H.S., Drăgan, L., Moreau, L.: An auditable reputation service for collective adaptive systems. In: Social Collective Intelligence. Springer (2014) 159–184
13. Markovic, M., Edwards, P., Corsar, D.: Sc-prov: A provenance vocabulary for social computation. In: International Provenance and Annotation Workshop, Springer (2014) 285–287
14. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
15. Cohen, M.H.: The unknown and the unknowable-managing sustained uncertainty. Western Journal of Nursing Research **15**(1) (1993) 77–96
16. Pearl, J.: Causality: Models, Reasoning and Inference. 2nd edn. Cambridge University Press, New York, NY, USA (2009)
17. Rubin, D.B.: Inference and missing data. Biometrika **63**(3) (1976) 581–592
18. Frey, B.J., Kschischang, F.R., Loeliger, H.A., Wiberg, N.: Factor graphs and algorithms. In: Proceedings of the Annual Allerton Conference on Communication Control and Computing. Volume 35., University of Illinois (1997) 666–680
19. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
20. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Trans. Inf. Theor. **47**(2) (September 2006) 498–519
21. Holleran, E., Bredahl, M.E., Zaibet, L.: Private incentives for adopting food safety and quality assurance. Food Policy **24** (1999) 669–683
22. Organization, W.H.: Risk assessments of Salmonella in eggs and broiler chickens. Volume 2. Food & Agriculture Organization (2002)